Statistics and Data Analysis

Distributions and Sampling (1)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Introduction

- We have learned two separate topics.
  - Descriptive statistics: visualization and summarization of **existing data** to understand the data.
  - Probability: using **assumed probability distributions** (for, e.g., inventory management).
- Now it is time to connect them.
- This lecture:
  - We will study how to **estimate the distribution** of a random variable from existing data.
  - We will study how to **sample** from a population.
- The next lecture:
  - We will study **sampling distribution**: the distribution of a sample.

Road map

- **Estimating probability distributions**.
  - When the sample space is small.
  - When the sample space is large.
- Sampling techniques.

Estimating probability distributions

- Given a random variable, how to know its **probability distribution**?
  - Given a coin, what will be the outcome of tossing it?
  - Given a room and a time point, what will be the temperature?
  - Given a population of people, what will be the age of a randomly selected person?
  - Given a potential customer, will she/he buy my product?
  - Given a web page and a time horizon, how many visitors will we have?
  - Given a batch of products, how many will pass a given quality standard?
- We want more than one value; we want a **distribution**.
  - For each possible value, how likely it will be realized.
  - We may plan our inventory level only if we have a demand distribution.
- To do the estimation, we **do experiments** or **collect past data**.

Estimating probability distributions

- Given a random variable, how to know its probability distribution?
  - Given a random variable X, how to get $F(x) = \Pr(X \leq x)$?
- Given a coin, how to know whether it is fair?
  - Let X be the outcome of tossing a coin.
  - Let $X = 1$ if the outcome is a head or 0 otherwise.
  - Let $\Pr(X = 1) = p = 1 - \Pr(X = 0)$.
  - Is $p = 0.5$?

Frequency and probability distributions

- The most straightforward way: Use a **frequency distribution** to be the **probability distribution**.
  - We may flip the coin for 100 times.
  - Suppose we see 46 heads and 54 tails.
  - We may "estimate" that $p = 0.46$.
- A frequency distribution and a probability distribution are different.
  - A frequency distribution is what we observe. It is an outcome of investigating a **sample**.
  - A probability distribution is what governs the random variable. It is a property of a **population**.
- We may never know whether we are right. Technically speaking, we will never be "right."
  - However, this is the most practical way.
  - This is "approximately right" if we have enough data.
  - "To what degree we are wrong" will be discussed in further lectures.

Estimating a discrete distribution

- Consider a discrete random variable whose number of possible values are not too many.
  - Tossing a coin: 2 possible values. Rolling a dice: 6 possible values.
  - The gender of a randomly selected student: 2 (or more) possible values.
  - The district that a randomly selected Taipei resident lives in: 12.
  - Tomorrow's weather situation: sunny, cloudy, raining, snowing.
  - The daily sales quantity of cars at the small car dealer: 0, 1, ..., 10.
- Let X be the random variable and S be the sample space.
  - We are saying that S does not contain too many values.
- We want to know $\Pr(X = x) = p_x$ for any $x \in S$.
- In this case, let $\{x_i\}_{i=1,...,n}$ be our observed sample data. Given a value $x \in S$, we will simply use the **proportion**

$$\frac{\text{number of } x_i\text{s that is } x}{\text{number of } x_i\text{s}}$$

to be our estimated p_x.

When the sample space is small: example

- A data set records the daily weather for the 731 days in two years.
  - 1 for sunny or partly cloudy, 2 for misty and cloudy, 3 for light snow or light rain, and 4 for heavy snow or thunderstorm.
- Let X be the daily weather for any given day in the future.
- We have $S = \{1, 2, 3, 4\}$.
- By looking at the data set, we obtain

x	Frequency	Proportion
1	463	0.633
2	247	0.338
3	21	0.029
4	0	0

- Let $p_i = \Pr(X = i)$, we then estimate that $p_1 = 0.633$, $p_2 = 0.338$, $p_3 = 0.029$, and $p_4 = 0$.

Manually adjusting an estimation

- The estimated probability distribution of X is

$$p_1 = 0.633,\ p_2 = 0.338,\ p_3 = 0.029,\ \text{and } p_4 = 0.$$

- We know that this estimation is just based on a sample.
  - It is never "right."
  - Manual adjustments based on experiences or knowledge are allowed.
- E.g., we may adjust it to

$$p_1 = 0.65,\ p_2 = 0.3,\ p_3 = 0.03,\ \text{and } p_4 = 0.02.$$

Refining an estimation

- The estimated probability distribution of X is

$$p_1 = 0.633,\ p_2 = 0.338,\ p_3 = 0.029,\ \text{and } p_4 = 0.$$

- We may refine the estimation by considering **more information**.
- Suppose that we know the day of interest is on December.
  - For the 62 days in December in our sample, we have

x	Frequency	Proportion
1	32	0.516
2	27	0.436
3	3	0.048
4	0	0

- We may adjust it (again with manual adjustments) to

$$p_1 = 0.5,\ p_2 = 0.4,\ p_3 = 0.06,\ \text{and } p_4 = 0.04.$$

When the sample space is large

- When the sample space is large, this method is not very helpful.
  - E.g., a data set records the daily bike rentals in 731 days.
  - Let X be the daily bike rental.
  - X is discrete. Its sample space contains more than 8000 values.
  - The naive counting for frequencies does not help.
- In this case, we rely on **frequency distributions** to estimate the probability for the value to be **within a class**.
  - We may use the **class midpoint** to represent values in the class.
  - We may generate a **uniform distribution** for each class.

When the sample space is large: example

- Let X be the daily bike rental for a given day in the future.
- A data set contains the daily bike rentals in 731 days.
- We obtain the frequency distribution of daily bike rentals:

x	Frequency	Proportion
$[0, 1000)$	18	0.025
$[1000, 2000)$	80	0.109
$[2000, 3000)$	74	0.101
$[3000, 4000)$	107	0.146
$[4000, 5000)$	166	0.227
$[5000, 6000)$	106	0.145
$[6000, 7000)$	86	0.118
$[7000, 8000)$	82	0.112
$[8000, 9000)$	12	0.016

Using class midpoints as representatives

- We now create an artificial sample space $S = \{500, 1500, ..., 8500\}$.
- We estimate that $\Pr(X = 500) = 0.025$, $\Pr(X = 1500) = 0.109$, ..., and $\Pr(X = 8500) = 0.016$.
- This probability distribution can help us predict daily bike rentals in the future.
- We may of course manually adjust or refine the estimated probabilities.

x	Proportion
$[0, 1000)$	0.025
$[1000, 2000)$	0.109
$[2000, 3000)$	0.101
$[3000, 4000)$	0.146
$[4000, 5000)$	0.227
$[5000, 6000)$	0.145
$[6000, 7000)$	0.118
$[7000, 8000)$	0.112
$[8000, 9000)$	0.016

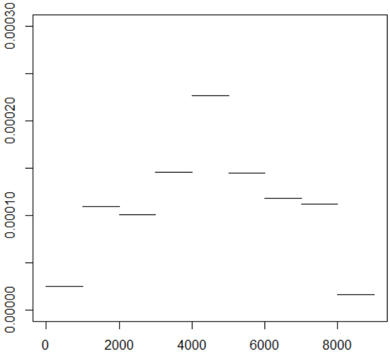
Generating uniform distributions for classes

- For each class, we create a uniform distribution so that its total probability is the observed proportion.
- Let $f(x)$ be the pdf of X for $x \in [0, 9000)$.
- Within $[0, 1000)$, the area below the pdf should be 0.025. This implies that $f(x) = \frac{0.025}{1000} = 0.000025$ for $x \in [0, 1000)$.
- Similarly, we have $f(x) = 0.000109$ for $x \in [1000, 2000)$.
- We repeat this process to all classes.

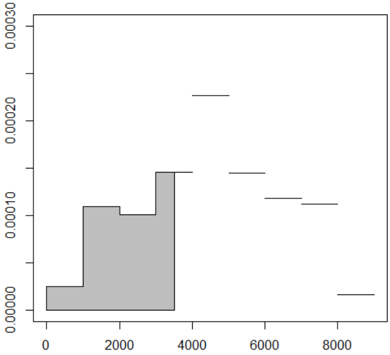
x	Proportion
$[0, 1000)$	0.025
$[1000, 2000)$	0.109
$[2000, 3000)$	0.101
$[3000, 4000)$	0.146
$[4000, 5000)$	0.227
$[5000, 6000)$	0.145
$[6000, 7000)$	0.118
$[7000, 8000)$	0.112
$[8000, 9000)$	0.016

Generating uniform distributions for classes

- The pdf $f(x)$ can be depicted:

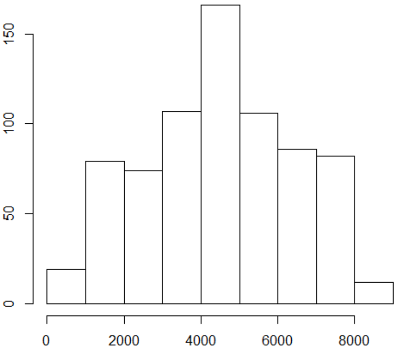


- The cdf $F(x)$ can be constructed:



Estimating a continuous random variable

- A continuous random variable "is" a discrete random variable with **extremely many** possible values in the sample space.
  - E.g., it is common in practice to approximate the daily bike rentals as a continuous random variable.
- We still start from a frequency distribution.
- The **histogram** now suggests us a continuous distribution.
  - Naturally, it looks similar to the pdf made by generating uniform distributions.

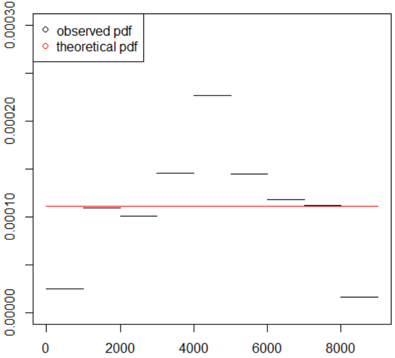


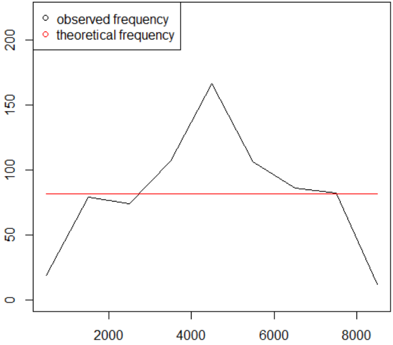
Fitting a distribution to a histogram

- We want to **fit** a distribution to a histogram.
- To do so, we select a distribution (by investigation and some experiences), find the theoretical frequency for each class following the distribution, and then plot the two sequences of frequencies together.
  - **Observed frequencies** are from the histogram.
  - **Theoretical frequencies** are from the assumed distribution.
  - If the two sequences are "close to each other," the fitting is appropriate.
- Equivalently, we may draw the pdf of the assumed distribution and the discrete distribution made by multiple uniform distributions together.
- We may try a few assumed distributions and select the best one.

Fitting a uniform distribution to a histogram

- Consider the daily bike rental example again.
- If we assume $X \sim \text{Uni}(0, 9000)$, we have $f(x) = \frac{1}{9000}$ for $x \in [0, 9000]$.
  - Or the theoretical frequencies are all $\frac{731}{9}$ in all classes.





- X does not seem to be $\text{Uni}(0, 9000)$.

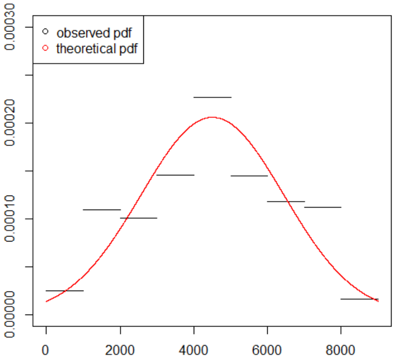
Fitting a normal distribution to a histogram

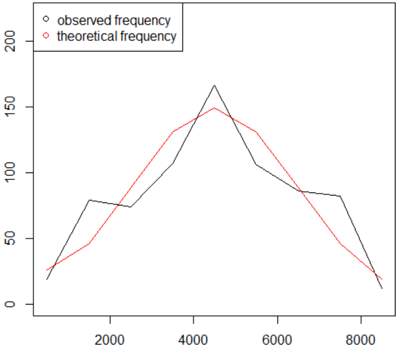
- Let's try to fit a normal distribution to the histogram.
- We need to choose a mean and a standard deviation to construct the normal curve.
  - People may use their judgment.
  - A typical way: Use the sample mean and sample standard deviation.
  - For the 731 values, we have $\bar{x} \approx 4504$ and $s \approx 1937$.
- Let's fit $\text{ND}(4504, 1937)$ to the histogram.
- If $X \sim \text{ND}(4504, 1937)$, we have:

$[l, u)$	$\Pr(l \leq X < u)$	Theoretical frequency
$[0, 1000)$	0.035	25.75
$[1000, 2000)$	0.063	45.92
	$\vdots$	
$[8000, 9000)$	0.025	18.59

Fitting a normal distribution to a histogram

- If we assume $X \sim \text{ND}(4504, 1937)$:





- $\text{ND}(4504, 1937)$ seems to fit the observed data better.
- Further trials and adjustments are always possible.

Summary

- We want to estimate the probability distribution of a random variable.
- When the sample space is small:
  - Use the relative frequency of each possible value to be its probability.
- When the sample space is large:
  - Construct a frequency distribution.
  - Use the relative frequency of each class to be its probability.
  - In each class, either put all the probability on the class midpoint or spread it to all values.
- When the sample space is extremely large:
  - Look at a histogram and guess which probability distribution fits it.
  - Find the theoretical frequency for each class.
  - Compare the two sequences of observed and theoretical frequencies.
  - Stop when the overall difference is "small."[1]
- Human judgments may be needed.

[1] How to define "small" will be discussed in further lectures.

Road map

- Estimating probability distributions.
- **Sampling techniques**.

Random vs. nonrandom sampling

- Sampling is the process of selecting a **subset** of entities from the whole population.
- Sampling can be **random** or **nonrandom**.
- If random, whether an entity is selected is **probabilistic**.
  - Randomly select 1000 phone numbers on the telephone book and then call them.
- If nonrandom, it is **deterministic**.
  - Ask all your classmates for their preferences on iOS/Android.
- Most statistical methods are **only** for random sampling.
- Some popular random sampling techniques:
  - Simple random sampling.
  - Stratified random sampling.
  - Cluster (or area) random sampling.

Simple random sampling

- In simple random sampling, each entity has **the same probability** of being selected.
- Each entity is assigned a label (from 1 to N). Then a sequence of n random numbers, each between 1 and N, are generated.
- One needs a **random number generator**.
  - E.g., `RAND()` and `RANDBETWEEN()` in MS Excel.

Simple random sampling

- Suppose we want to study all students graduated from NTU IM regarding the number of units they took before their graduation.
  - $N = 1000$.
  - For each student, whether she/he double majored, the year of graduation, and the number of units are recorded.

i	1	2	3	4	5	6	7	...	1000
Double major	Yes	No	No	No	Yes	No	No		Yes
Class	1997	1998	2002	1997	2006	2010	1997	...	2011
Unit	198	168	172	159	204	163	155		171

  - Suppose we want to sample $n = 200$ students.

Simple random sampling

- To run simple random sampling, we first generate a sequence of 200 random numbers:
  - Suppose they are 2, 198, 7, 268, 852, ..., 93, and 674.
  - Sampling with or without replacement?
- Then the corresponding 200 students will be sampled. Their information will then be collected.

i	1	**2**	3	4	5	6	**7**	...	1000
Double major	Yes	**No**	No	No	Yes	No	**No**		Yes
Class	1997	**1998**	2002	1997	2006	2010	**1997**	...	2011
Unit	198	**168**	172	159	204	163	**155**		171

- We may then calculate the sample mean, sample variance, etc.

Simple random sampling

- The good part of simple random sampling is **simple**.
- However, it may result in **nonrepresentative** samples.
- In simple random sampling, there are some possibilities that **too much** data we sample fall in **the same stratum**.
  - They have the same property.
  - For example, it is possible that all 200 students in our sample did not double major.
  - The sample is thus nonrepresentative.

Simple random sampling

- As another example, suppose we want to sample 1000 voters in Taiwan regarding their preferences on two candidates. If we use simple random sampling, what may happen?
  - It is possible that 65% of the 1000 voters are men while in Taiwan only around 51% voters are men.
  - It is possible that 40% of the 1000 voters are from Taipei while in Taiwan only around 28% voters live in Taipei.
- How to fix this problem?

Stratified random sampling

- We may apply **stratified random sampling**.
- We first split the whole population into several **strata**.
  - Data in **one** stratum should be (relatively) **homogeneous**.
  - Data in **different** strata should be (relatively) **heterogeneous**.
- We then use simple random sampling for each stratum.
- Suppose 100 students double majored, then we can split the whole population into two strata:

Stratum	Strata size
Double major	100
No double major	900

Stratified random sampling

- Now we want to sample 200 students.
- If we sample $200 \times \frac{100}{1000} = 20$ students from the double-major stratum and 180 ones from the other stratum, we have adopted **proportionate** stratified random sampling.

Stratum	Strata size	Number of samples
Double major	100	20
No double major	900	180

- If the opinions in some strata are more important, we may adopt **disproportionate** stratified random sampling.
  - E.g., opening a nuclear power station at a particular place.

Stratified random sampling

- We may further split the population into more strata.
  - Double major: Yes or no.
  - Class: 1994-1998, 1999-2003, 2004-2008, or 2009-2012.
  - This stratification makes sense **only if** students in different classes tend to take different numbers of units.
- Stratified random sampling is good in **reducing sample error**.
- But it can be hard to identify a reasonable stratification.
- It is also more **costly** and **time-consuming**.

Cluster (or area) random sampling

- Imagine that you are going to introduce a new product into all the retail stores in Taiwan.
- If the product is actually unpopular, an introduction with a large quantity will incur a huge lost.
- How to get an idea about the popularity?
- Typically we first try to introduce the product **in a small area**. We put the product on the shelves only in those stores in the specified area.
- This is the idea of **cluster (or area) random sampling**.
  - Those consumers in the area form a sample.

Cluster (or area) random sampling

- In stratified random sampling, we define strata.
- Similarly, in cluster random sampling, we define **clusters**.
- However, instead of doing simple random sampling in each strata, we will only choose **one or some clusters** and then collect **all** the data in these clusters.
  - If a cluster is too large, we may further split it into multiple **second-stage clusters**.
- Therefore, we want data in a cluster to be **heterogeneous**, and data across clusters somewhat **homogeneous**.

Cluster (or area) random sampling

- In practice, the main application of cluster random sampling is to understand the popularity of **new products**. Those chosen cities (counties, states, etc.) are called **test market cities** (counties, states, etc.).
- People use cluster random sampling in this case because of its feasibility and convenience.
- We should select test market cities whose population profiles are similar to that of the entire country.

Nonrandom sampling

- Sometimes we do **nonrandom sampling**.
- Convenience sampling.
  - The researcher sample data that are easy to sample.
- Judgment sampling.
  - The researcher decides who to ask or what data to collect.
- Quota sampling.
  - In each stratum, we use whatever method that is easy to fill the quota, a predetermined number of samples in the stratum.
- Snowball sampling.
  - Once we ask one person, we ask her/him to suggest others.
- Nonrandom sampling **cannot** be analyzed by the statistical methods we introduce in this course.